|  |
| --- |
| Crime Data from 2020 to present  Statistical Analysis |
| Laiba Khan Probability and Applied Statistics  CSCI-3327 Fall 2023 |



Table of Contents

**Introduction1**

About Me2

The Dataset3

**Problems4**

Chapter 15

Chapter 25

Chapter 35

Chapter 45

Chapter 55

**Solutions4**

**Analysis4**

**References4**

Introduction

About me

Hello. My name is Laiba Khan. I am a senior currently attending Stockton University. One thing you may know about me is that I am currently in the final days of my Probability and Applied Statistics class.

This class, although tough, made me learn a significant amount about professionalism and how a job may present work to its employees in a similar way that this course does. This specifically, was a project that a particularly enjoyed doing. Being able to apply real world data into programs and solving statistical problems was an interesting experience.

I will continue to pursue efforts in learning about the work-style of professional jobs and continue to put forth the most professional work that I can.

Thank you for reading.

The Dataset

Data.gov

This dataset was sourced from the official government database catalog. Data.gov is a platform or repository that offers a variety of real-world data for analysis and problem-solving. You can find diverse datasets covering a wide range of subjects, from government data to research findings and more.

It's a hub for data enthusiasts, researchers, and anyone interested in leveraging data for analysis or decision-making. The informal and user-friendly interface makes it easier for people to navigate and discover the datasets they need.

The Sampled Data

The data that was chosen from Data.gov is a collection of Los Angeles crime records dating back to 1/1/2020 12:00:00 AM. Overall, this dataset allows for the analysis of various aspects of reported crimes, including the types of crimes, their locations, and details about victims. Researchers, law enforcement, and the public can use this data to gain insights into crime patterns and trends in the Los Angeles area.

According to the site, the data is transcribed from original crime reports typed onto paper (\*\*\*\*). There may be some inaccuracies and missing data due to this, as some locations are not specified, etc. The dataset is also *very* large, including up to 857658 entries. The problems written within this report may simplify the data to be better suited.

Here is a guide to the labels used within the CSV file:

*These are inferenced, and this may not be 100% correct.*

|  |  |
| --- | --- |
| **DR\_NO** | Report number for the crime incident. |
| **Date Rptd** | Date when the crime was reported. |
| **DATE OCC** | Date when the crime occurred. |
| **TIME OCC** | Time when the crime occurred. |
| **AREA** | Numeric code representing the policing area. |
| **AREA NAME** | Name of the policing area. |
| **Rpt Dist No** | Report district number. |
| **Part 1-2** | Classification of the crime as Part 1 or Part 2. |
| **Crm Cd** | Numeric code representing the crime type. |
| **Crm Cd Desc** | Description of the crime. |
| **Mocodes** | Modus operandi codes for the crime. |
| **Vict Age** | Age of the victim. |
| **Vict Sex** | Gender of the victim. |
| **Vict Descent** | Descent (ethnicity) of the victim. |
| **Premis Cd** | Numeric code representing the location type. |
| **Premis Desc** | Description of the location. |
| **Weapon Used Cd** | Numeric code representing the weapon used (if any). |
| **Weapon Desc** | Description of the weapon used. |
| **Status** | Status of the crime report. |
| **Status Desc** | Description of the crime report status. |
| **Crm Cd 1-4** | Additional crime codes if applicable. |
| **LOCATION** | General location of the crime. |
| **Cross Street** | Cross street information. |
| **LAT** | Latitude coordinate of the crime location. |
| **LON** | Longitude coordinate of the crime location. |

Problems

Chapter 1: What is Statistics?

* 1. Introduction

For each of the following situations, identify the population of interest, the inferential objective, and how you might go about collecting a sample.

1. The US Government wants to estimate the proportion of crimes with women involved among crimes committed in Los Angeles during the current year.
2. Assess whether there is a predominant preference among adult residents for a certain time to commit crimes.
3. A crime-analyst wants to estimate the district numbers with the most crime.
   1. Graphical Methods

The most common ages of victims of crimes, ranked by oldest to youngest:

50 55 54 32 33 34 35 64 39 47 25 43 50 61

Construct a relative frequency histogram to describe these data.

Chapter 2

2.3 A Review of Set Notation

From a survey of 50 victims of crime in Los Angeles, it was found that 9 were from N Hollywood, 36 were men, and 3 were women from N Hollywood. Find the number of these students who were

1. Men, from N Hollywood, or both.
2. Women not from N Hollywood.

2.9 Calculating the Probability of an Even: The Event-Composition Method

Three patrol cars, operating independently, are set to crimes occuring through a certain area. Each set has a probability of .01 of failing to criminal in its area. If a crime occurs in the area, what is the probability that

1. a crime goes undetected?
2. is detected by all three radar sets?

2.10 The Law of Total Probability and Bayes Rule

When criminals are are reported, 90% of the individuals are accurately reported. Also, if it was a true report, the test will show that the report is correct with the probability .9. Only 1% of the population incorrectly reports crimes. If a person is chosen at random from the the number of reports and through further investigation it turns out the report was correct, what is the conditional probability that the report is in fact correct? Are you surprised by the answer? Would you call this test reliable?

2.11 Numerical Events and Random Variables

Refer to Exercise 2.120. Let the random variable Y represent the number of defective refrigerators found after three refrigerators have been tested. Compute the probabilities for each value of Y .

Chapter 3

3.2 The Probability Distribution for a Discrete Random Variable

Four crimes are known to have two falsely reported. An investigator tests the components one at a time until the two falsely reported are realized. Once she locates the two false reports, she stops testing, but the another investigator does a second investigation to ensure accuracy. Let Y denote the number of the investigation on which the second false-report is found. Find the probability distribution for Y.

3.3 The Expected Value of a Random Variable or a Function of a Random

Variable

Approximately 10% of the reports are reported in Central Los Angeles. If two crime reports are randomly selected, find the mean and variance of the number of crimes in Central Los Angeles.

There is 58055 entries with Central Los Angeles.

3.4 The Binomial Probability Distribution

In 2020, the amount of crime reports in Los Angeles that year was 199,490. Suppose that approximately 1% of all reports that year was of bike theft and that 100 reports are randomly selected from the dataset. Which of the following random variables has a distribution that can be approximated by a binomial distribution? Whenever possible, give the values for n and p.

1. The number of bike theft reports
2. The amount of time required by the police officer to fill out the report
3. The number of female victims in the sample

3.5 The Geometric Probaility Distribution

Suppose that 30% of the crimes committed towards identified women occur during the night. Reports are looked at sequentially and are selected at random from the pool. Find the probability that the first applicant that is a women and occurs during the evening is found on the fifth report.

3.7 The Hypergeometric Probability Distribution

A sample of reports contains 10 reports, 5 of which are of theft. A crime investigator selects five of the reports at random to investigate, not knowing which are for theft. What is the probability that all five of reports are for theft.

3.8 The Poisson Probability Distribution

Let Y be a random variable representing the number of crimes reported in a specific area of Los Angeles during a certain time period, with an average rate (*λ*) of 3 crimes per unit of time.

3.11 Tchebysheff’s Theorem

Let Y be a random variable representing the number of crimes reported in a specific area of Los Angeles during a certain time period, with a mean *μ* of 11 crimes and a variance of 9.

Find a lower bound for P(6 < Y < 16).

Chapter 4

4.3 Expected Value for Continuous Random Variables

The rate of reports taken in 2020 has a density function is given by f (y) = {1/3, 100 ≤ y ≤ 25000, 0, elsewhere, indicating the acceptable range for reports need to be taken., elsewhere. Find E(Y) and V(Y). Do not evaluate V(Y) fully, give me the expression to solve it as it is too complex.

4.4 Uniform Probability Distribution

If a law enforcement agent responds to incidents and lands at a random point on a line between markers A and B, find the probability that the agent is closer to A than to B. Additionally, determine the probability that the agent's distance to A is more than three times the distance to B, representing the likelihood of responding to incidents in specific areas.

4.5 The Normal Probability Distribution

The width of criminal activity hotspots is normally distributed with a mean of 950 units (e.g., incidents) and a standard deviation of 10 units.

1. What is the probability that a randomly chosen hotspot has activity levels between 947 and 958 units?
2. What is the appropriate value for C such that a randomly chosen hotspot has activity levels less than C with probability 0.8531?

4.11 Expectations of Discontinuous Functions and Mixed Probability

Distributions

The length of time required to complete a college achievement test is found to be normally distributed with mean 70 minutes and standard deviation 12 minutes. When should the test be terminated if we wish to allow sufficient time for 90% of the students to complete the test?

Chapter 5

5.2 Bivariate and Multivariate Probability Distribution

Of nine executives in a business firm, four are married, three have never married, and two are divorced. Three of the executives are to be selected for promotion. Let Y1 denote the number of married executives and Y2 denote the number of never-married executives among the three selected for promotion. Assuming that the three are randomly selected from the nine available, find the joint probability function of Y1 and Y2.

5.3 Marginal and Conditional Probability Distributions

In Exercise 5.1, we determined that the joint distribution of Y1, the number of contracts awarded to firm A, and Y2, the number of contracts awarded to firm B, is given by the entries in the following table. y1 y2 012 0 1/9 2/9 1/9 1 2/9 2/9 0 2 1/90 0

a Find the marginal probability distribution of Y1.

b According to results in Chapter 4, Y1 has a binomial distribution with n = 2 and p = 1/3. Is there any conflict between this result and the answer you provided in part (a)?

Solutions

* 1. Introduction

1. The population of interest is women who have been part of a crime. The inferential objective is to estimate the proportion of crimes with women involved in crimes committed in Los Angeles during the current year. You may use a stratified method. This is because you can use them to divide up subgroups within your data, based on features like gender or identity.
2. The population of interest is reported crimes. The inferential objective is to estimate difference in amount of crime during certain times of day. You may collect a sample with cluster random sampling. With that it selects random sampling throughout different times of day where the measure the amount of reported crimes. Then you observe and compare with enough data.
3. The Population of interest is all reported case of vandalism made in Central. The inferential objective is to find the average rate of vandalism in central area based on a sample. This can provide insight into the broader pattern of crime and property damage in this area. You may collect a sample using a random sampling method. You can start at a random day and time and select every nth reported case of vandalism. It will represent a diverse set of vandalism cases with different times and circumstances while also telling the amount of reports.
   1. Graphical Methods

A graph of a number of people

Description automatically generated

The first value (<=0) can be chalked up to a lack of input or a possible reference to unborn children. It is not clear in the data, but it may be one of the two.

Chapter 2

2.3 A Review of Set Notation

1. = 14(Women) – 9 (From N Hollywood)

= 5

* 1. Calculating the Probability of an Even: The Event-Composition Method

1. a crime goes undetected?

=0.03

1. is detected by all three radar sets?

=0.00002619

* 1. The Law of Total Probability and Bayes Rule

P(correctly reporting | Investigations shows correctness) = P(A|B)

= 0.988

= 98.8% correct. This is an effective test, but the context determines whether or not the results are reliable.

Chapter 3

3.2 The Probability Distribution for a Discrete Random Variable

P(Y=2) =

P(Y=3) =

3.3 The Expected Value of a Random Variable or a Function of a Random

Variable

= 2 \* 0.1 = 0.2

= 2 \* 0.1 \* (1-0.1) = 0.18

n = 2, p = 0.1

the mean is 0.2 and variance is 0.18

3.4 The Binomial Probability Distribution

The first option (a), “the number of bike theft reports” has a distribution that can be approximated. N would be the number of reports randomly selected and p would be the probability of it being a bike theft report.

3.5 The Geometric Probability Distribution

Suppose that 30% of the crimes committed towards identified women occur during the night. Reports are looked at sequentially and are selected at random from the pool. Find the probability that the first applicant that is a women and occurs during the evening is found on the fifth report.

= 0.07203

3.6 The Negative Binomial Probability Distribution

Refer to Exercise 3.90. If each test costs $20, find the expected value and variance of the total cost of conducting the tests necessary to locate the three positives.

3.7 The Hypergeometric Probability Distribution

3.8 The Poisson Probability Distribution

1. .1680

.1493 + 0.2240 + 0.2240 = 0.5973

3.11 Tchebysheff’s Theorum

= .64

Chapter 4

4.3 Expected Value for Continuous Random Variables

- > too complex to evaluate

4.4 Uniform Probability Distribution

If a law enforcement agent responds to incidents and lands at a random point on a line between markers A and B, find the probability that the agent is closer to A than to B. Additionally, determine the probability that the agent's distance to A is more than three times the distance to B, representing the likelihood of responding to incidents in specific areas.

4.5 The Normal Probability Distribution

The width of criminal activity hotspots is normally distributed with a mean of 950 units (e.g., incidents) and a standard deviation of 10 units.

1. What is the probability that a randomly chosen hotspot has activity levels between 947 and 958 units?
2. What is the appropriate value for C such that a randomly chosen hotspot has activity levels less than C with probability 0.8531?

Chapter 5

5.2 Bivariate and Multivariate Probability Distribution

Of nine executives in a business firm, four are married, three have never married, and two are divorced. Three of the executives are to be selected for promotion. Let Y1 denote the number of married executives and Y2 denote the number of never-married executives among the three selected for promotion. Assuming that the three are randomly selected from the nine available, find the joint probability function of Y1 and Y2.

5.3 Marginal and Conditional Probability Distributions

In Exercise 5.1, we determined that the joint distribution of Y1, the number of contracts awarded to firm A, and Y2, the number of contracts awarded to firm B, is given by the entries in the following table. y1 y2 012 0 1/9 2/9 1/9 1 2/9 2/9 0 2 1/90 0

a Find the marginal probability distribution of Y1.

b According to results in Chapter 4, Y1 has a binomial distribution with n = 2 and p = 1/3. Is there any conflict between this result and the answer you provided in part (a)?

Analysis

References

U.S. Government. (Year). Crime Data from 2020 to Present [Dataset]. Data.gov. <https://catalog.data.gov/dataset/crime-data-from-2020-to-present>